tagtog: Interactive Human and Machine Annotation of Gene Mentions in PLOS Full-Text Articles

Juan Miguel Cejuela, Peter McQuilton¹, Laura Ponting¹, Steven J. Marygold¹, Raymund Stefancsik¹, Gillian H. Millburn¹, Burkhard Rost² and the FlyBase Consortium⁺.

Paul-Hindemith-Allee 6, 80939, Munich, Germany; ¹Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK; Bioinformatics-i12, ²Department of Informatics, Technical University of Munich (TUM), 85748 Garching, Germany;

To whom correspondence should be addressed: pam51@gen.cam.ac.uk


Abstract
We present the tagtog system, a web-based annotation framework that can be used to mark-up biological entities and concepts in full-text articles. tagtog leverages user manual annotations in combination with automatic machine-learned annotations to provide accurate gene symbol and name identification in biomedical literature. For this submission we present, in collaboration with the FlyBase database curation team at Cambridge University, the task of identifying and extracting mentions of Drosophila melanogaster gene symbols and names in full-text biomedical articles from the PLOS stable of journals. We show here the results of three experiments with different sized corpora, and assess gene recognition performance. Finally, we would like to extend an invitation for Biocurators at the BioCreative IV Track 5 -- User Interactive Task (IAT) to come and find us to try tagtog themselves.

Introduction
tagtog (http://tagtog.net) is a web-based framework for the annotation of named entities. A user creates a project, defines a named-entity recognition task (such as gene annotation), and uploads a set of text documents to the system. Each document is then displayed in a web editor where the user can add, delete, or correct the information relevant to the annotation task. An example of the user interface is shown in Figure 1. The user can add the annotation of an entity by selecting the corresponding word(s) and remove it by clicking another time on the selection. During the
course of their work, a biocurator will need to analyse thousands or even millions of documents, an undertaking that is impossible to do through manual means alone. To address this problem, the tagtog system leverages machine-learning methods to perform the same type of annotations computationally. Initially, the tool is trained with a small set of manually annotated documents. Once trained, tagtog can be used to process a set of novel documents wherein automatically generated predictions are made that then can be reviewed and corrected by the user. It is this continuous and interactive re-training of the machine learning methods with user feedback, that can lead to an ever-improving performance in automatic prediction. Once optimized, the trained machine learning methods can be used to process and annotate a large volume of documents to a sufficiently accurate level. We envisage this process will save curator time and effort and lead to significant increases in gene annotation efficiency. Finally, the annotated documents can be exported (in anndoc XML format) for the particular user's application needs, such as for further curation or manipulation into curation records for database submission.

In the following section, we describe the system's current features and technical details. In the final section, we will describe our three experiments to date, using papers from the FlyBase (http://flybase.org) bibliography.

**Current Features**

- **Browser support**: The system runs on all major current browsers only requiring HTML5 and javascript. Chrome and Firefox are officially supported. Other browsers like Opera, Safari, and Internet Explorer (9 and 10) are regularly tested but lack official support. Access to the system can be provided using our invitation system.
– **Multiple projects**: Users can create different annotation projects and load their own dictionaries and corpora.

– **Multiple users**: Multiple users on the same project are also supported, allowing curation teams to view and annotate the same set of papers.

– **Multiple entities**: Support for the annotation of multi-entities in the same project is also desired and expected to be ready by the end of 2013.

– **Active learning**: tagtog actively asks for user feedback on predicted annotations. A proposed mechanism was already developed in an early version of tagtog, presented last year at the BioCreative 2012 workshop.

– **Document searching**: Papers can be searched using the search tool at the top of the interface. Options include searching by document ID (based on the DOI), entities, or whether a paper has been fully annotated or not. In the future we hope to add the facility to search by PubMed ID (PMID).

– **Import options**: Any paper following the NCBI Journal Publishing Tag Set\(^1\) or the BioMed Central format\(^2\) can be uploaded to tagtog. This includes full-text papers from the PLOS, BioMed Central, Chemistry Central, and Springer Open collections. In the near future, we will accept papers from the new JATS format\(^3\) and plain text files.

– **Export options**: Annotations for each paper can exported as a tab-separated list of terms linked to PMIDs. In addition, *anndoc* XML for entire corpora can be downloaded, allowing users to archive annotations. We hope to allow soon export of annotations in the new BioC format\(^5\).

**Defining the Annotation Guidelines**

Upon project creation, the first step for a user is to define the annotation guidelines (Figure 2). These define what and how to annotate the project documents. The user currently has the following options:

– **Entity**: choose the entity class name to annotate.

– **Entity Dictionary**: upload a user-defined dictionary/ontology of collected entity names. The dictionary can contain synonyms and database-specific IDs, allowing data integrity checks and seamless integration of the results with the parent database.

– **Meta Information**: define a list of checkboxes for document triage, e.g., whether the article contain disease mentions, or information on a new transgene.

– **Annotables**: select the sections of the full-text articles that can be annotated (and trained with). The annotation of figures' and images' captions is decided independently: *always*, *never*, or *section-dependent*.

– **Pre-Annotations**: if activated, upon entity selection or de-selection, all instances of an annotation will be ‘pre-annotated’ or automatically de-selected. These pre-annotations require user confirmation before final annotation.
**Defining a Corpus**

All full-text XML documents that follow the NCBI Journal Publishing Tag Set format (versions 2.x and 3.0) (for example all PLOS journals) and the BioMed Central format (for example, all articles from BioMed Central, Chemistry Central, and SpringerOpen) can be uploaded either as single documents or in batches. The system's internal parser recognizes the documents' sections, subsections, figures, tables, and some additional meta-information such as the paper's original URL. The project corpus can be augmented progressively as the user sees fit. Currently, documents are placed in two folders, the *pool* folder, where training documents are placed, and the *gold* folder, where a smaller set of previously annotated documents are used for the evaluation of the machine learning methods' performance (these documents are never used for training).

**Downloading Your Data**

The user can export the entire corpus upon request. Documents are annotated in XML documents with an in-house-defined format, called *anndoc*. The current version of anndoc (0.3) supports in-line entity annotations (without being nested), meta information, hierarchy of sections, figures, tables, and their captions. Anndoc follows a HTML-like format, with a header for meta
information and a body for the document's text and annotations. The tags have been so selected to match those of HTML5's for an easy display of the document in current browsers. In addition, the annotations for each document can also be exported as a tab-separated list of terms linked to the corresponding PMID. In the future, we are keen to adopt other standard output formats if there is sufficient user demand.

The Machine Learning Component of tagtog
A core, defining characteristic of the system is that the users can choose the entity type to annotate, such as gene, GO term, or disease. The system boasts a general-purpose named-entity recognizer. The recognizer is customized to the prediction task at hand by means of user feedback and by the dictionary/ontology of entity terms. The system is configured to allow expansion of annotation types with new machine annotators via plugins to enable annotation of multiple data-types and languages.

If desired, the machine-learning component of tagtog can be turned off to allow biocurators to use the tagtog interface exclusively for manual curation.

Interactive Annotation Task

FlyBase Gene Mention Annotation Task
In collaboration with the FlyBase database (http://flybase.org) at Cambridge University, we have undertaken the task of identifying and extracting mentions of genes of the drosophila genus (fruit flies) in full-text biomedical articles. Since September 2012, FlyBase have been testing tagtog for its adoption in their Genetic Literature Curation pipeline. Currently, FlyBase has two well-defined application cases for tagtog:

- **Gene to publication links**: To use tagtog to automatically generate gene-publication links. These will be used to populate the FlyBase gene report and reference report web pages, as well as to pre-populate the community curation FlyBase Fast-Track-Your-Paper tool and FlyBase curation records.
- **Skim Curation**: In addition to the gene-to-publication links mentioned above, FlyBase curators could use the metadata information tags to triage publications for further curation. This would replicate the ‘skim curation’ step in the FlyBase paper curation pipeline.

In this task, we will show how FlyBase could integrate tagtog with their curation pipeline, address specific practical details such as how to consume the generated output formats, what is the subjective user experience with the tool on a daily basis, and what short-comings and improvements can be identified. In addition, during this process we will expand the FlyBase-tagtog corpus with more manually annotated documents and assess the performance of the machine learning methods. We hope to release this corpus for use by other text-mining groups as
we believe the corpus to be the largest and most complete gene mention annotation set in full-text articles currently available.

To date, FlyBase curators have manually annotated 431 papers using the tagtog interface. All papers are full-text PLOS journal articles from between 2011 and 2013. The following document sections were annotated: title, abstract, results, materials and methods, and figure and table legends. The paper annotations have been used to iteratively train the machine-learning component of tagtog.

So far, we have performed three annotation+benchmark iterations. In the first two iterations, annotations were done partially manually by a sole curator (Peter McQuilton) and automatically by the system. In the third iteration, all five FlyBase curators annotated papers manually. All the manual annotations and corrections were performed using tagtog's document editor interface.

The document sets for the three iterations are as follows:

- **Iteration 1**: A sole curator (P. McQuilton) manually annotated a training set of 20 articles. Trained with these documents, the system was applied to predict an unlabelled test set of 99 articles. The curator then went through the test set and, corrected, added, or removed the predicted annotations when appropriate. In the end, mismatched annotations between the original predictions and the revised annotations were counted as errors.

- **Iteration 2**: the previous two sets were united to form a training set of 119 articles. For evaluation, the user manually annotated a test set of 20 articles (which we will refer to as the ‘Gold Standard’. The system was trained on the 119 articles and benchmarked against the 20 test articles. In contrast to Iteration 1, in this case prediction errors could be read off directly from mismatches against the test set.

- **Iteration 3**: FlyBase jamboree combined set. The previous two sets, plus a further 312 papers curated by 5 different FlyBase curators, were combined to form an annotated corpus of 431 Fly-related papers. These papers were used to train tagtog before assessment on the Gold Standard set.

For the performance benchmarks, we used standard evaluation measures for named-entity recognition (NER), namely: precision (P), recall (R), and F1-Measure (F1). Only exact matches between the predictions and the test annotations were counted as correct. That is, the predictions had to match the same exact word boundaries. Two types of counts were considered: 1) unique entities on a document basis. That is, for a test entity X in a document, the predictions were right if at least one mention of that entity could be identified in that document, wrong otherwise. Equivalently, all unique entities identified by the predictions but not present on the test annotations were counted as errors. 2) All entity mentions for all documents. That is, for all entity mentions, matching predictions and test annotations were counted as correct. Mismatched mentions, either false positives or false negatives, were counted as errors. Note that for testing, only the annotable sections defined by the curator are compared.
Figure 3 shows the entity recognition performance for all entity mentions in a paper, that is, the ability of tagtog to identify the presence of a gene mention, either as a symbol or name. The figure shows that the performance has steadily improved in proportion to the corpus size. The same performance improvement behavior is seen for unique entity recognition (figure 4), that is, the ability to identify the presence of a gene mention at least once in a paper.

**Figure 3** - Entity recognition performance over all three corpora sizes. Iteration 1 was using a corpus of 20 documents to train tagtog to identify gene symbols/names in 99 documents. Iteration 2 used a training set of 119 documents to assess performance on a ‘gold standard’ set of 20 papers. Iteration 3 used the FlyBase corpus, composed of 431 articles, to train tagtog, with performance assessed on the gold standard set.
Figure 4 – Entity instance recognition performance over all three corpora sizes. Iteration 1 was using a corpus of 20 documents to train tagtog to identify gene symbols/names in 99 documents. Iteration 2 used a training set of 119 documents to assess performance on a ‘gold standard’ set of 20 papers. Iteration 3 used the FlyBase corpus, composed of 431 articles, to train tagtog, with performance assessed on the gold standard set.

BioCreative IAT challenge
In addition to the experiments mentioned above, we are recruiting other biocurators from outside FlyBase to participate in the testing and assessment of tagtog. In this challenge, biocurators will be first asked to manually annotate mentions of an entity class in up to 20 papers using tagtog, and then asked to annotate a second set of 20 papers that have been marked up by the machine-learning component of tagtog, trained on the first set of 20 papers. Note that the entity class to
annotate will be defined by the biocurators. Both annotation tasks will be timed, enabling judgement not only of the accuracy of the tagtog annotation but also the time cost/benefit of these assisted annotations.

Conclusions
Although overall a moderate performance, we believe these early evaluation results to be promising. To our knowledge, these results represent one of the first NER evaluations with a substantial amount of full-text articles in the biomedical field. NER with full-text articles is understood to be considerably more difficult than for abstracts, a focus scope that was more studied in the past\textsuperscript{7,8}. It must be noted that the machine learning method is not specialized to the problem, except for the ontology of terms given by the user. In particular, we have not specialized the machine learning to focus on the genes of a single organism, \textit{Drosophila} melanogaster. On the whole, prediction performance appears to increase with an increase in the volume of training data. The continuous learning of tagtog is designed to generate cheaper (in terms of manual curation effort) training data, by taking advantage of semi-automatic annotation. Still, a central goal for the curation task proposed in this submission, is to further improve the performance of the tagtog system so it operates at a sufficiently accurate level that it can be incorporated into the FlyBase literature curation pipeline.

Author Contributions
JMC and PM devised the experiments and wrote the manuscript. PM led the curation process and gave directions as for the development of the user interface. PM, SM, LP, RS, and GM annotated the corpus. JMC develops tagtog.

Funding
This work was supported by private funding for Juan Miguel Cejuela and the National Human Genome Research Institute at the National Institutes of Health (P41 HG00739), and the Medical Research Council (UK) (G1000968) for members of FlyBase.

References
2. BioMed Central format. URL http://www.biomedcentral.com/about/xml.
3. PLOS. URL http://www.plos.org/.